(*Mis*)*computation* in Computational Psychiatry

Matteo Colombo
https://mteocolphi.wordpress.com/
m.colombo @ uvt .nl

Abstract An adequate explication of *miscomputation* should do justice to relevant practices in the computational sciences. While philosophers of computation have neglected scientific practices outside computer science, here I focus on computational psychiatry. I argue that computational psychiatrists use a concept of *miscomputation* in their explanations, and that this concept should be explicated as interest-relative and perspectival, although non-arbitrary, relatively clear-cut, experimentally evaluable, and instrumentally useful. To the extent my argument is convincing, we should reconsider the general adequacy of the mechanistic view of computation for illuminating relevant methodological and explanatory practices in the computational sciences.

Keywords: miscomputation; computational psychiatry; aberrant prediction error; aberrant precision; malfunction; representation

1. Introduction

Because computing systems are kinds of rule-governed systems, they can perform computations *wrong*. A computing system, that is, can return an output o_2 that deviates to a greater or a lesser extent from the output of the function f on input i, $f(i) = o_1$, which the system ought to return. When this happens, the system *miscomputes*.

Philosophers of computation have explicated the concept of *miscomputation* without paying much attention to relevant scientific practices outside computer science (Fresco & Primiero 2013; Dewhurst, 2014; Piccinini 2015; Tucker 2018). In this paper, I extend this line of work on *mis-*

computation to computational psychiatry, and address these two questions: Does a concept of *miscomputation* have any place in computational psychiatry? If it does, how should it be explicated?

My answer to the first question is that a concept of *miscomputation* figures at least in Bayesian and Reinforcement Learning computational modelling practices in psychiatry. Psychiatrists often use this concept for explaining impairments associated with psychiatric illnesses. These explanations involve expressions like "malfunctioning computations," "false inference," "aberrant prediction error" or "aberrant precision estimates," which are plausibly associated with the concept of *performing a computation wrong*, as opposed to performing different kinds of computations.

My answer to the second question is that this concept of *miscomputation* should be understood as interest-relative and perspectival, although non-arbitrary, relatively clear-cut, experimentally evaluable, and instrumentally useful. If any concept of *computation* entails the concept of *miscomputation*, then at least one adequate explication of *computation* should also be interest-relative and perspectival.

To be clear: my focus, here, is not on whether brains are objectively physical computing systems, or whether they must have representational properties if they actually are computers. My focus is on certain scientific practices, imputations and interpretations. My overall point is that a purely mechanistic notion of *miscomputation* does not fit some imputations, interpretations and practices central to computational psychiatry. This meta-scientific conclusion is meant to put pressure on the idea that a mechanistic explication of *miscomputation* suffices to do justice to relevant practices involved in the computational sciences.

I begin by outlining the aims and methodologies of contemporary computational psychiatrists, showing that the concept of *miscomputation*

has a place in psychiatry and that *miscomputation* cannot be chalked off as indicating only a difference in computing (Section 2). After I lay out two possible explications of *miscomputation* (Section 3), I argue that a satisfactory explication of *miscomputation* in computational psychiatry should refer to psychiatrists' expectations and pragmatic concerns in relation to the (mal)functioning and representational properties of a target system modelled as a computational system. I develop this argument based on the idea that computational psychiatrists rely on *specifications* of target systems (Section 4). In a short conclusion, I summarize the contribution of this paper, and draw one implication for the mechanistic view of computation.

2. *Miscomputation* in computational psychiatry

Computational psychiatrists use computer simulation, computational and mathematical modelling, and computational methods for pursuing the goals of classification, diagnosis, prediction, understanding, and treatment (e.g., Ahmed, Graupner, & Gutkin 2009; Huys, Moutoussis, & Williams 2011; Montague, Dolan, Friston, & Dayan 2012; Deco & Kringelbach 2014; Friston, Stephan, Montague, & Dolan 2014; Adams, Huys, & Roiser 2016; Kurth-Nelson et al. 2016; Brugger & Broome 2019).

To pursue these goals, there are theory-driven and data-driven approaches. Typical in data-driven approaches is the use of machine-learning techniques to mine large sets of genetic, neural and behavioural data from psychiatric patients and healthy controls, for patterns, clusters, and causal dependencies (Huys, Maia & Frank 2016, 405-8). Theory-driven approaches generally seek to assess people's performance in experimental tasks, to evaluate the effectiveness of therapies, and to explain psychiatric phenomena by imputing mathematical functions to be computed to experimental participants or target neural systems, and by modelling the activi-

ties and components of these systems in terms of computations of these functions (e.g., Maia & Frank 2011).

Computational psychiatrists need not be committed to the idea that neural systems are actual computing systems to successfully pursue their goals. Computational psychiatrists may or may not believe that the brain is actually a computing system, or that it is in some sense an information-processing system. But this does not matter to the success of their modelling practices.

Like in other fields in the sciences of mind and brain, the emphasis is on successful computational modelling. On successfully representing target systems in terms of rule-governed transitions from mathematical inputs to mathematical outputs (Egan 2019). This requires that researchers fit computational models to various sets of data, and generate simulation data from the best fitting model to ensure the model is empirically adequate. Given a set of candidate models for a clinically relevant phenomenon, the most empirically adequate model will be the most useful to pursue the goals of classification, diagnosis, explanation, or treatment with respect to that phenomenon.

Let me describe a typical study in computational psychiatry, which illustrates this point. Schlagenhauf and collaborators (2014) wanted to explain why patients diagnosed with schizophrenia show an impairment in certain learning tasks. Using a model-based brain imaging methodology (e.g., Colombo 2014a), they collected behavioural and neural data from un-medicated patients diagnosed with schizophrenia and healthy controls. Their experimental participants performed a probabilistic reversal learning task,[1] while undergoing magnetic resonance brain imaging.

[1] This task requires participants to learn from probabilistic feedback, where the structure of the task can change so that what used to be positive outcomes (i.e., a

Schlagenhauf and collaborators formulated various computational models corresponding to different hypotheses about the rule-governed transitions from inputs to output, which could describe participants' behaviour in their task. They evaluated the empirical adequacy of these competing models based on individual participants' trial-by-trial choice and neural data. One model had the best fit to data from both healthy controls, and only some schizophrenia patients. For most schizophrenia patients, the best fitting model was a different one.[2]

Schlagenhauf and collaborators identified strong associations between the activity of target neural systems in individual participants and trial-by-trial variation in specific components of the best fitting models. For all participants, activity in the ventral striatum in response to the same patterns of state-reward contingencies in the learning task was most strongly associated with a component of the models called "reward prediction error"—more on this component in Section 4 below. Compared to healthy controls, all schizophrenia patients exhibited reduced activity in the ventral striatum. Schizophrenia patients, whose choice and neural data were captured by the same model as in the healthy controls, showed a level of prefrontal activity similar to that of healthy controls, but higher than that in the other patients. Overall, both reduced activity in the striatum and in

---

positive reward) are now negative outcomes (i.e., a punishment, or negative reward), and what used to be negative are now positive outcomes.

[2] Specifically, the best fitting model for healthy controls and some schizophrenia patients was a Hidden Markov Model. According to this model, participants built and updated a representation of the structure of the task, based on the past history of choices and resulting rewards. Their belief about the current state of the task would be used to make a choice. Instead, the best fitting model for the other schizophrenia patients was a Rescorla-Wagner model. According to this model, participants did not build a representation of the structure of the task. For each trial, participants would choose an option based on its expected value. After a trial, the expected value of only the chosen option would be updated on the basis of a prediction error (Schlagenhauf et al. 2014, 172-3).

the prefrontal cortex of participants correlated with higher scores of positive symptoms of schizophrenia such as delusions and hallucinations assessed with the Positive and Negative Syndrome Scale (PANSS) (Kay, Fiszbein, & Opler 1987).

From these findings, Schlagenhauf et al. (2014) concluded two things. First, reduced reward prediction error signals in the ventral striatum is a general dysfunction in schizophrenia—even when the performance of both schizophrenia patients and healthy controls is captured by the same type of computational model. Second, reduced reward prediction error signals in the ventral striatum explains schizophrenia patients' impaired performance in reversal learning tasks—even when we control for differences in computational ascriptions to different sub-groups of patients.

Although Schlagenhauf et al. (2014) did not mention the term "miscomputation," they framed their conclusions in terms of a "dysfunction" consisting in the "impairment" (172) or "deficit" of "ventral striatum prediction error signaling" (178). This way of talking is plausibly associated with the idea of *performing a computation wrong*, as opposed to performing different kinds of computations, or implementing different kinds of computational architecture. After all, Schlagenhauf et al. (2014) relied on computational modelling exactly to reach "more definitive conclusions... about processes more directly related to the disease that diminishes problems of interpretation due to behavioural differences associated with adaptive disease dependent strategies" (172).

Several other studies can be cited to show that the concept of *miscomputation* has a place in computational psychiatry, and that this concept cannot be understood only in terms of differences in computations, or differences in computational architecture.

In the context of Bayesian and Reinforcement Learning modelling (cf., Colombo 2019), Montague et al. (2012) are explicit that computational psychiatrists seek "to characterize mental dysfunction in terms of *aberrant computations*" (72, emphasis added). Even more explicit are King-Casas et al. (2008), when they write that computational modelling "offers the opportunity to understand some of the components of [psychiatric] disorders in terms of *malfunctioning computations*" (806, emphasis added). Huys, Guitart-Masip, Dolan & Dayan (2015) distinguish three classes of "failure modes" that computational modelling uncovers in mental illnesses, namely: performing the right computations to solve the wrong problem, performing *poor* or *wrong computations* to solve the right problem, and performing the right computations to solve the right problem but in an unfortunate environment (for example, an environment that makes generalization from experience more difficult or maladaptive).

Two prominent examples of Bayesian and Reinforcement Learning miscomputations concern *prediction error* and *precision estimates*. A prediction error is a component of many Bayesian and Reinforcement Learning models, which quantifies the difference between an expected outcome and the actual outcome—for example, the difference between the expected monetary value of making a choice and the actual amount of money received in making that choice. Precision estimates of an outcome are components of many Bayesian models, and quantify the inverse variance of the outcome—for example, they are estimates of how far a set of monetary gains is spread out from the mean monetary gain in the set and from one another.

Schlagenhauf et al. (2014) refer to prediction error computations in a Reinforcement Learning model when they conclude that schizophrenia patients have a dysfunction in ventral striatum reward prediction error sig-

nalling. Fletcher & Frith (2009) also refer to prediction error computations when they suggest that psychotic symptoms of schizophrenia, such as hallucinatory experiences and delusional beliefs, can usefully be explained "in terms of a disturbed hierarchical Bayesian framework," specifically in terms of a disruption in prediction error signalling (48). Examining autonomic arousal and cortical activity in patients with autism spectrum disorder (ASD), Gu et al. (2015) refer to precision estimates to interpret their findings. They say: "the current findings provide direct support for recent proposals suggesting that failures in Bayesian inference, and particularly aberrant precision (i.e., inverse variance) of the information encoded at various levels of sensorimotor hierarchies, may contribute to socioemotional deficits in ASD" (3335). Lawson et al. (2018) also talk about precision estimates in their study of learning in autistic patients. Testing the computational prediction that aberrant precision estimates explain autistic patients' reduced behavioural surprise to atypical events, they conclude that their "findings provide preliminary empirical evidence for neurobiologically informed Bayesian accounts of autism that emphasize... *inappropriate* setting of gain (precision) on cortical responses (prediction errors) under conditions of uncertainty" (1298).

This overview highlights two aspects of contemporary practice in computational psychiatry. First, computational psychiatrists use terms like "aberrant prediction error" and "aberrant precision estimates" for explaining psychiatric phenomena. For example, aberrant prediction error computations would explain schizophrenia patients' impairment in reversal learning, as well as their hallucinatory experiences and delusional beliefs. Aberrant computations of precision estimates would explain autistic patients' socioemotional deficits, as well as their impaired responses to environmental volatility. Second, when computational psychiatrists say that

prediction error signaling is aberrant, what they plausibly mean is not that the target of their best-fitting computational model is functioning a-typically, or in a statistically abnormal way. What they mean is that it malfunctions, or presents some dysfunction.

Given these two aspects of existing practice in computational psychiatry, one may ask a number of questions. Specifically, how should we exactly understand terms like "aberrant prediction error" or "aberrant precision estimates"? What is a good thing to mean by these terms in the specific context of Bayesian and Reinforcement Learning modelling for the purpose of explaining clinically relevant phenomena? (on the idea of an explication as a "good thing to mean" see Gupta 2015 Sec. 1.5).

3. Two explications of *miscomputation*

Here are two possible answers to these questions:

[*m-miscomputation*] A target system miscomputes just in case the best-fitting computational model of the system captures some malfunction in the system, where (i) the system's malfunctioning is determined in relation to the system's objective goals or selective history, and where (ii) the ascription that the system (mis)computes a certain function in a task does not presuppose any representational ascription to the system.

[*p-miscomputation*] A target system miscomputes just in case the best-fitting computational model of the system captures some malfunction in the system, where (i') the system's malfunctioning is determined in relation to certain expectations, interests and conventions of a relevant scientific community, and where (ii') the ascription that the system

(mis)computes a certain function in a task presupposes representational ascriptions to the system.

Both explications are committed to the idea that the condition that is both necessary and sufficient to apply the concept of *miscomputation* in psychiatry is that a computational model successfully represent some malfunction of a target system in a task. As explained in Section 2, the criterion of success here is fitting the experimental data into computational models that predict the data itself. So, for example, we can say that ventral striatum prediction error signalling counts as a miscomputation if and only if the best-fitting model of striatal activity in a task posits computations of prediction errors, these posits predict relevant neural and choice data generated by striatal activity sufficiently well, and the striatum is somehow malfunctioning.

The two explications differ in their commitment as to whether or not a system's malfunctioning is determined objectively just on the basis of mind-independent properties of the system, and in their commitment as to whether or not ascribing that the system (mis)computes a certain function in a task presupposes representational ascriptions to the system.

The first explication has much in common with prominent accounts of concrete computing systems in the mechanistic tradition, such as Piccinini's (2015). I call it *m-miscomputation*. The second explication is the conjunction of a perspectival view about the function of performing computations in a task (e.g., Dewhurst 2018b) and a pragmatist view about representation (e.g., Egan 2010; Egan 2014; Coelho Mollo 2020). I call it *p-miscomputation*.

To unpack the commitments of *m-miscomputation* and *p-miscomputation*, it will help to rehearse relevant ideas from the literature

on concrete computation. I start from the mechanistic account of concrete computation, and focus on various treatments of (mal)function. Then, I briefly review three popular accounts of how representational content is determined.

3.1 On malfunction

According to the mechanistic account, concrete computing systems are mechanisms that perform computations, that is: systems of spatially and temporally organized, causally related components with functions to perform. At least one function of computing mechanisms is that of performing computations (Miłkowski 2013; Fresco 2014; Piccinini 2015; Coelho Mollo 2019).

There are at least three options about what determines the function to compute of a mechanism. According to the first option, the function to compute of a mechanism is determined by the stable causal contributions that performing this function makes in relation to some objective goal, where the objective goals of an organism are its survival and inclusive fitness (cf., Maley & Piccinini 2017).

The second option is that the function to compute of a mechanism is determined by the stable causal contributions that performing this function made, in the past, to processes of differential reproduction and differential retention (e.g., processes of evolution, development, and learning) involving organisms with that type of mechanism in a population (cf., Neander 1991; Garson 2019).

While the second option says that a mechanism's function to compute depends on the selective history of the mechanism, the first option does not appeal to any historical process, but only to how a mechanism's

performing computations contributes, now, to the survival and inclusive fitness of organisms with that kind of mechanism.

However, the first and second option are analogous because they share the idea that what fixes the function to compute of a mechanism are objective, mind-independent properties of the mechanism. An explication of *miscomputation* committed to this idea will recommend that the ascription that the brain's computational function in a given task is, say, to compute posterior probabilities, or to map situations to actions so as to maximize some measure of reward, should be understood independently from any human interest or expectation. These computational functions would amount to biological functions. Their ascriptions to human brains would be warranted to the extent we have warranted beliefs that computing posterior probabilities (or maximising some specific measure of reward), now, furthers the objective goals of humans; or that computing posterior probabilities (or maximizing some specific measure of reward), in the past, causally contributed to the differential retention of a brain with certain features in humans within a population.

According to a third option, the function to compute of a mechanism is partly determined by certain expectations, interests and conventions of a relevant scientific community. In particular, Dewhurst (2018b, 581) argues that it is determined by certain interpretations of the physical structure of the mechanism, where these interpretations are grounded in an "explanatory perspective." An explication of *miscomputation* committed to this idea will say that the meaning of the ascription that the brain's computational function is to compute posterior probabilities is dependent on certain expectations, explanatory interests, and conventions of some relevant community. Computational functions would not just amount to biological functions. Ascriptions of certain computational functions to target systems

in a task would be warranted to the extent relevant, perspectival interpretations of structural and causal features of the target system are warranted.

Now that we have a better idea of how a mechanism's function to compute can be determined, let's consider *mal*function. In the context of artificial computers, Piccinini (2015, 149-50) claims that miscomputation is a "failure of a hardware component to perform its function." This failure can be caused by some (non-essential)[3] component of the system being missing, or by the alteration of the spatial, temporal or causal organization of the hardware. Regardless of how it is caused, a hardware component's failure to perform its computational function consists in a deviation between the function the component should compute and what the component actually computes.[4] That is, the system "*M* is computing function *f* on input *i*, $f(i) = o_1$, *M* outputs o_2, and $o_2 \neq o_1$" (Piccinini 2015, 13).[5] Fresco and Primiero (2013) call this deviation "operational error," and Turing (1950, 449) calls it "error of functioning."[6] Depending on the right option

[3] If an essential component of a computing system is missing, altered or broken, then the system may not compute anymore. If a system does not compute at all, then it cannot miscompute.

[4] There's no consensus among proponents of the mechanistic view about how we should individuate what a computing system actually computes at a time. For example, unlike Piccinini (2015), Tucker (2018, 8) argues that a system's computational structure is individuated without any reference to factors external to the system; what the system is actually computing at a time is determined by the actual inputs to the system at that time, in addition to its computational structure.

[5] In Section 2, I referred to Huys, Guitart-Masip, Dolan & Dayan (2015), who distinguished three classes of "failure modes" that computational modelling highlights in mental illnesses. One failure mode, *viz.* performing the right computations to solve the wrong problem, arises when the system M returns o_2, while computing a function $g(i)$, which differs altogether from the $f(i)$ it ought to compute. In this case, o_2 may be the right output to solve the wrong problem, $g(i)$.

[6] Writes Turing: "We may call [… these two types of errors] 'errors of functioning' and 'errors of conclusion'. Errors of functioning are due to some mechanical or electrical fault which causes the machine to behave otherwise than it was designed to do. In philosophical discussions one likes to ignore the possibility of

about what determines a mechanism's function to compute, there are three ways to articulate the nature of this deviation, and, thereby, the nature of computational malfunction.

First option: when a system computes function *f* on input *i*, the system returns output o_2; o_2 deviates from the output $f(i) = o_1$; and o_1 would make, now, a causal contribution to some objective goal of the system.

Second option: when a system computes function *f* on input *i*, the system returns output o_2; o_2 deviates from the output $f(i) = o_1$; and o_1 made a causal contribution, in the past, to processes of differential reproduction and differential retention for some trait.

Third option: when a system computes function *f* on input *i*, the system returns output o_2; o_2 deviates from the output $f(i) = o_1$; and o_1 is the output a relevant community expects for systems of that type, given a certain "explanatory perspective," interests, and conventions.

The first and second way to articulate computational malfunction are reflected in *m-miscomputation*. If an adequate explication of *miscomputation* reflects either of these two options, then warranted ascriptions that a brain is malfunctioning in a given task when it computes, say, posterior probabilities depends on warranted beliefs that its output o_2 deviates from that output o_1, which either furthers the objective goal of the organism, or causally contributed to the differential retention of brains in a certain population of organisms. Instead, if an adequate explication of *mis-*

---

such errors; one is therefore discussing 'abstract machines'. These abstract machines are mathematical fictions rather than physical objects. By definition they are incapable of errors of functioning. In this sense we can truly say that 'machines can never make mistakes'. Errors of conclusion can only arise when some meaning is attached to the output signals from the machine. […] When a false proposition is typed we say that the machine has committed an error of conclusion. There is clearly no reason at all for saying that a machine cannot make this kind of mistake." (Turing 1950, 449).

computation reflects the third way of articulating the idea of computational malfunction, then warranted ascriptions that the brain is malfunctioning when it computes posterior probabilities would depend on warranted, communal expectations about outputs o_2 and o_1, given certain pragmatic interests and conventions.

3.2 On representation

Unlike *m-miscomputation*, *p-miscomputation* is committed to the idea that (ii') (mis)computation in a task should presuppose representational ascriptions. This idea is reflected in the semantic view of concrete computation, according to which a system cannot compute unless it possesses representational properties (e.g., Fodor 1975; Churchland & Sejnowski 1992; Sprevak 2010; Rescorla 2014; Shagrir 2018). According to this view, computing systems differ from non-computing systems because computing systems can manipulate representations, while non-computing systems cannot.

It is plausible that the *individuation* of systems that compute does not involve any representation. After all, a machine can systematically manipulate strings of digits, following a rule defined over the appropriate degrees of freedom of its possible input strings, outputs and internal states, even if the strings have no representational property (see, e.g., Dewhurst 2018a).[7]

Yet, in the computational sciences, representation plays several fruitful roles. For example, some computer scientists and engineers design and build certain machines to execute appropriate mathematical computa-

[7] By 'degrees of freedom', I mean one of two things: either certain formal syntactic differences, or certain concrete physical differences between inputs and outputs and states of a system along some dimension of variation (e.g., voltage levels, rate of activation, or timing of activation).

tions. They, and anybody else, describe these machines as doing maths. But it is only by presupposing that the states of these machines represent numbers that these descriptions and practices make sense. So, even if the semantic view of concrete computation is false, it remains an interesting question what practices and ascriptions in the computational sciences presuppose the ascription of representational properties to a system, and what purposes these ascriptions could serve.

To evaluate the role of representational ascriptions in relation to *miscomputation* in computational psychiatry, it will help to briefly rehearse different proposals about how the content of a representation gets fixed—that is, how the condition for a representation's being right (or wrong) about a subject matter is determined.

Three proposals are prominent in the existing literature. According to the first proposal, the contents of a system's representations are determined, narrowly, by the system's intrinsic properties. The idea is that the content of a subject's representation does not require the subject to stand in any relation to anything in the environment. The contents of our thoughts would depend only on the causal goings-on inside our heads (cf., Fodor 1987). The condition for a representation's being right about a subject matter would be an intrinsic property of our brains. If this condition is fulfilled, that representation is accurate (or true).

If content is determined narrowly, then computing systems with the same intrinsic properties must have the same representations. In the context of computational modelling in psychiatry, this proposal invites the prediction that modellers ascribe representations to target systems without appealing to features of the systems' environment, focusing only on features intrinsic to the systems.

According to the second proposal, the contents of a system's representations are determined, widely, by relevant extrinsic properties of the system. The idea is that the content of a subject's representation depends on the way the subject is embedded in the environment. Thus, the contents of our thoughts would depend both on the internal interactions between various states of our brain, as well as their relations to external circumstances. A brain state would represent the presence of a green tree in the environment because of some causal, information, historical or biological relation with green trees in the outside world (cf., e.g., Dretske 1981; Millikan 1984). The condition for a representation's being right about a subject matter would be an extrinsic property of our brains; it would involve the external condition required for the behavioural effects, which the representation prompts, to achieve certain ends. If this condition is fulfilled, that representation is accurate (or true).

If content is determined widely, then computing systems with the same intrinsic properties, but embedded in different social or physical environments, need not have the same representations. In the context of computational modelling in psychiatry, this proposal invites the prediction that modellers ascribe representations to target systems by appealing to features of the systems' environment, focusing on stable relations between features intrinsic to the systems and conditions in the world.

According to the third proposal, the content of a representation is fixed in a perspective-dependent fashion, or as Shagrir (2018) puts it "interpretatively." The idea is that the contents of a subject's representations are not objective properties, either narrow or wide. Although statements involving representations aim to state certain facts, they do not aim at truth. Because they aim at serving pragmatic purposes of a certain community—such as classification, prediction, explanation and intervention—

these statements should be accepted if they actually serve these purposes (cf., Dennett 1987; Egan 2014; Sprevak 2013).

If content is determined interpretatively and pragmatically in this way, then computing systems with the same intrinsic properties and embedded in the same social and physical environments need not have the same representations. In the context of computational modelling in psychiatry, this proposal invites the prediction that modellers ascribe representations to target systems pragmatically and interpretatively, based on the extent to which these ascriptions serve their purposes.

4. Explicating (*mis*)*computation* in computational psychiatry

Piccinini (2015) claims that "miscomputation finds an adequate explication within the mechanistic account" (275). In this section, I examine whether this claim is true in the context of Bayesian and Reinforcement Learning approaches in computational psychiatry. I use Schlagenhauf et al.'s (2014) study introduced above as a case study, and address these questions: When researchers say that a system's performing aberrant prediction error computations explains a certain psychiatric phenomenon, what is it that warrants their ascriptions of aberrant prediction error computations in a given task? What is it that warrants the idea that the system is malfunctioning? Is it some of the researchers' pragmatic interests, conventions and warranted "perspective"? Or, is it their warranted beliefs about the selective history or objective goals of the system? And should the ascription that the system (mis)computes prediction errors in a given task presuppose any representational ascription to the system?

4.1 Perspectival malfunction

Let's start from malfunction. Schlagenhauf et al. (2014) wanted to better understand why schizophrenia patients show an impairment in reversal learning tasks. The most successful behaviour in these tasks can be defined as the behaviour that maximises rewards, where rewards may consist in money, food, water, or some other good participants would find rewarding. Accordingly, one's behaviour is successful in this task to the extent it brings about specific rewarding outcomes.

Maximising rewards (and minimising losses) in reversal learning tasks depends on various capacities. One is the capacity to learn the state-reward contingencies in the task from experience. Another is the capacity of converting beliefs about the reward values into choices. Yet another one is the capacity to inhibit actions that are learned in response to certain cues when they no longer result in reward. These capacities can work more or less well. For example, learning can be more or less quick, the motivation to pursue subjectively rewarding outcomes can be more or less strong, or the inhibition of learned actions can be more or less effective. Where these capacities are impaired, participants in a reversal learning task will be less likely to flexibly change their behaviour in response to changes in the structure of the task, and so, less likely to maximise rewards and minimize losses in the task.

From behavioural, neural, and computational modelling results, Schlagenhauf et al. (2014) concluded that a dysfunction in prediction error computations in the ventral striatum could explain schizophrenia patients' impaired reversal learning. This dysfunction would explain why schizophrenia patients' behaviour is less successful in this task compared to healthy participants.

According to *m-miscomputation*, the ascription of a dysfunction in prediction error signalling in the ventral striatum means that, in schizo-

phrenia patients, either dopamine-dependent activity in the striatum does not return the outputs it was selected to return in reversal learning tasks, or it does not return those outputs that would promote schizophrenic patients' objective goals of survival and reproduction when they face these tasks.

This explication does not do justice to relevant practices. For two reasons. Call the first reason "the critical range problem." The problem is that an adequate explication of *miscomputation* should make sense of how and why computational psychiatrists often conclude that *reduced* or *increased* prediction error signalling in the ventral striatum is a dysfunction.

To illustrate the problem, suppose that some particular response activity in the ventral striatum is widespread among the participants in reversal learning tasks, but some smaller groups of participants exhibit reduced (or increased) activation.

If we accept *m-miscomputation*, then we need three premises to license the conclusion that ventral striatal prediction error computing is dysfunctional in the subgroups of participants. First, one has to map features of the task faced by the participants onto features of some real-world environment, with which humans recurrently interacted, or interact now. Second, one has to map participants' ventral striatal activations in this task onto ventral striatal activations in response to some matching real-world environment, with which humans recurrently interacted, or interact now. And finally, one has to show that given these mappings, a specific range of ventral striatal activation in response to reversal learning tasks was adaptive, or is adaptive now, and activations outside that range were likely, or are likely, to impede one's chance of survival and reproduction. If either of these premises is unwarranted, then the ascription that reduced (or increased) ventral prediction error computing is dysfunctional is unwarranted too.

Although researchers could rely on various types of evidence—e.g., ecological data, genetic data, phylogenetic data, comparative data—to warrant those premises, we have so far very little knowledge about *a critical range* of dopamine turnover in the ventral striatum for adaptive reversal learning (cf., Alcaro, Huber, & Panksepp 2007; O'Connell & Hofmann 2011; Howes & Kapur 2009). So, *m-computation* is currently of little help to explicate in what sense *reduced* or *increased* prediction error signalling in the ventral striatum counts as a dysfunction for computational psychiatrists.

The second reason why *m-miscomputation* is not a good thing to mean by expressions like "dysfunction of prediction error signalling" concerns the "mismatch problem." The problem is that an adequate explication should capture normal psychiatric usage of the term "dysfunction" in the face of possible *mismatches* between the computational function ascribed to a system and the environment with which the system would now compute that function. Let me explain.

Suppose that certain patterns of activations in the human dopamine system in response to certain physiological or environmental conditions are selected effects—one possible example might be the pattern of activation underlying the formation of certain beliefs in response to very surprising perceptual experiences. Based on *m-miscomputation*, we would consider those patterns to be a biological function of the dopamine system. Suppose that prediction error signals within a certain range in certain computational models in a given task show a good degree of fit with those patterns exhibited, now, by patients with delusions diagnosed with schizophrenia. We would then be warranted to say that computing certain prediction errors is (probably) a biological function of those dopamine responses. Suppose finally that there is an evolutionary mismatch between

the way the dopamine system is designed to respond to surprising perceptual experiences, and the response that would be adaptive with respect to the perceptual experiences in the current environment (Pani 2000). On *m-miscomputation*, one would not be warranted to say that those patterns of dopamine activity are dysfunctional, though they are statistically abnormal and are now associated with delusions exhibited by patients with schizophrenia. They would be functional responses of the dopamine system, which may produce delusions associated with schizophrenia given the current (mismatched) perceptual environment (cf., Garson 2019, 180-1).

Let's grant that existing evidence warrants this kind mismatch, and that the pattern of dopamine activation underlying the formation of certain beliefs in response to surprising perceptual experiences is a selected effect. One problem with *m-miscomputation* is that its recommendations go against normal psychiatric judgement. If the patterns of activation exhibited by schizophrenia patients are both mismatched and functional, then conclusions like the one drawn by Schlagenhauf et al. (2016) that reduced prediction error signals in the ventral striatum is a "signature dysfunction" of schizophrenia are false; we should not take them seriously. It would also be wrong to say that "that *dysfunction* of the mesocorticolimbic dopamine system causes delusion formation via *disrupted* prediction-error signalling" (Corlett et al. 2007, 2387-8, emphases added; see also Feeney et al. 2017).

If these conclusions are false, then one practical consequence is that interventions targeting changes in dopamine activity in schizophrenia would be misguided and potentially bad for patients. Because these interventions are often effective and have contributed to elucidate common characteristics of the pathophysiology of schizophrenia patients (Tsou

2012), understanding expressions like “dysfunctional striatal prediction errors” in terms of *m-miscomputation* would be practically unfruitful too.

P-miscomputation provides us with a better explication, which can make good sense of both the critical range problem and the mismatch problem. Both problems can be addressed if we understand ascriptions of computational (mal)function in a task as dependent on pragmatically useful representational ascriptions and a relevant explanatory perspective.

Let’s start from the idea of an explanatory perspective. In the context of computational psychiatry, this idea can helpfully be understood by analogy with *specifications* in computer science (Turner 2011; Fresco & Primiero 2013).

Specifications of a computational system are sets of documented, explicit requirements at various levels of abstraction, which a computer should satisfy. Specifications stipulatively define the vehicles of computing of a system (e.g., voltages, electric currents) and their rules of transformation, given the relevant degrees of freedom of a concrete physical system. Since specifications could be used to fabricate computers, and to evaluate their performance in a given task along various dimensions (e.g., processing power, energy consumption, memory, scalability, sturdiness), they function as blueprints and reference documents for computer scientists, engineers, programmers, computer manufacturers and users. They also enable consistent, transparent communication about a certain type of system.

Most importantly, they provide us with *stipulative definitions* of when and to what extent computing machines malfunction. As Turner puts it: “it is the act of taking a definition to have normative force over the construction of an artefact that turns a mere definition into a specification... Whether a [computational system] malfunctions is then not a property of

the [system] itself but is determined by its specification" (Turner 2011, 140-1). Or, in the words of Schweizer (2019, 41): "[i]t is only at a *non-intrinsic* prescriptive level of description that 'breakdowns' can occur, and we characterize these phenomena as malfunctions only because our extrinsic ascription has been violated."

Computational psychiatrists' explanatory perspectives can helpfully be understood by analogy with computer scientists' specifications. Such perspectives warrant "extrinsic ascriptions" that the range of activity exhibited by a certain neural system modelled as a computing system in a task is (dys)functional, or that the activity exhibited by that system in certain populations in a certain environment is plausibly dysfunctional, even though it may be an adaptation.

A computational framework like Reinforcement Learning is an example of a specification, which provides researchers with an explanatory perspective, or explanatory template, for studying and understanding the behaviour of certain biological and artificial systems, and of psychiatric phenomena too (Sutton & Barto 2018; Niv 2009; Maia & Frank 2011).

P-miscomputation handles the critical range problem by saying it is computational psychiatrists' explanatory perspective or specification that can warrant their ascription that a certain range of prediction error computation counts as a malfunction. When psychiatrists model ventral striatal activity in terms of prediction error signals, warranted claims about what range of the mathematical function returning prediction errors is dysfunctional and what range indexes well-functioning computing in various experimental participants depend on three sources of information belonging to their explanatory perspective. First, on optimality results in mathematics and computer science; second, on known associations between various profiles of prediction error signalling exhibited by different groups

participants in different experimental tasks; third, on diagnostic information about participants' general levels of suffering and "adaptive functioning" outside the lab (e.g, participants' PANSS scores).

Claims of (sub)optimality depend on mathematical results and on computer simulations. These results demonstrate under what conditions (e.g., under what parametrizations, in problems with what statistical or topological structure) a given Reinforcement Learning model quickly, and with little energy expenditure, can converge to a global (or local) maximum (or minimum) value of a function to be computed. These results set a normative standard, a yardstick, against which the learning performance of biological or artificial systems can be evaluated (Sutton & Barto 2018; Niv 2009).

Apart from results about optimality and computational complexity, the kind of specifications shared by computational psychiatrists can be related to individual and group differences in general levels of adaptive functioning and symptom severity. Computational psychiatrists form warranted expectations about these differences based on their clinical experience, calibrated scales like PANSS, and on widely shared diagnostic manuals like the DSM-5 and ICD-10, which define adaptive functioning in terms of "how well a person meets community standards of personal independence and social responsibility, in comparison to others of similar age and sociocultural background" (DSM-5, 31).

Now, computational psychiatrists sometimes find that some neural systems of some groups of psychiatric patients can adequately be modelled as performing optimal computations, or computations that are more efficient, or more accurate than the computations ascribed to healthy individuals to solve the same task—patients with depression, for example, show an absence of unrealistic optimism, which may captured with optimal compu-

tations in some tasks (cf., Huys, Daw, & Dayan 2015). And yet, psychiatrists understand these optimal computations as *miscomputation*, either because, based on results from computer science and mathematics, these optimal operations are known to involve trade-offs in efficiency, reliability and timeliness with other computations in other tasks, or because, based on clinical experience and diagnostic information, these optimal computations are known to be associated with low levels of adaptive functioning or with some debilitating symptom.

Recall that Schlagenhauf et al. (2014) found that the range of magnitudes of prediction error signalling in healthy controls was larger than in schizophrenia patients, who displayed reduced prediction error signals in the ventral striatum. Computer simulations show that reduced prediction error signals lead to blunted updates of the expected values of outcomes in a given state for future trials, which means that learning becomes slower and worse than learning driven by relatively higher prediction errors. So, the outputs returned by the dopamine system of schizophrenia patients diverged from the outputs a dopamine system ought to return, where this "ought" is grounded in a communal specification (or explanatory perspective) of the dopamine system as a reinforcement learning computing system, and on warranted expectations based on clinical experience and shared tools for diagnosis.

In summary, the "right" (or "wrong") range for the values of prediction error signals is based on "extrinsic ascriptions," on a communal specification. Such ascriptions are non-arbitrary, because they are based on reproducible and transparent optimality results and on communal expectations about certain illnesses. They are relatively clear-cut, because they give us determinate answers for many profiles of prediction error signalling. They are experimentally evaluable and revisable in the light of new

optimality results, accumulating clinical experience, and revisions of widely shared diagnostic tools. They are instrumentally useful too, since psychiatrists can use these perspectival ascriptions of dysfunction for classification and devising targeted therapies (cf. Colombo & Heinz 2019 on classifications based on computational phenotypes).

4.2 Pragmatist representation

Let's finally consider the role of representation in ascriptions of miscomputation in a task. Unlike *m-miscomputation*, *p-miscomputation* invites us to understand these ascriptions by positing representations. But what epistemic or practical role could representations play here exactly?

As I noted at the beginning of this section, Schlagenhauf et al. (2014) started with a cognitive task, *viz*. with a reversal learning task, where schizophrenia patients show an impairment. Successful performance in this task can be defined in terms of the relationships between participants and the environment, *viz*. as participants' interactions with the environment that maximize their rewards. Defining the task and the performance to be explained in this way involves representational ascriptions to participants. For example, it involves the ascription that participants have beliefs and expectations about reward contingencies in the task, the desire to obtain as much reward as possible, or the ability of using their beliefs and desires to make choices.

Schlagenhauf et al. (2014) adopted the explanatory perspective of Reinforcement Learning to explain participants' performance in this task. The Reinforcement Learning models they formulated need not involve any representational posit. Prediction error signals in these models quantify the difference between the learned predictive value of some current state and the sum of the current reward and the value of the next state. Specifically,

a reward prediction error signal $\delta(t)$ computed at time t is equal to $r(t) + V(t+1) - V(t)$, where $V(t)$ is the predicted value of some option at time t, and $r(t)$ is the reward outcome obtained at time t. Because *any* distal state could in principle bear predictive value, Reinforcement Learning models compute prediction errors *regardless* of the environment they would find themselves in. In this specific sense, they are environment-neutral (or domain general). This means that the models Schlagenhauf et al. (2014) formulated would compute the same mathematical functions $V(t)$ and $\delta(t)$ over certain inputs, had the input states been strings of sounds instead of the geometrical shapes Schlagenhauf et al. (2014) actually used to distinguish different states in their learning task.

Representational ascriptions played the role of connecting ascriptions of Reinforcement Learning (mis)computations with the behaviours exhibited by experimental participants in a given task. In order to clarify how their computational modelling results explained performance in reversal learning, and, in particular, how aberrant prediction error signals explained impaired performance in patients with schizophrenia, Schlagenhauf et al. (2014) interpreted operations and components of computational models in terms of representations of specific states and reward outcomes in their task. And these interpretations allowed them to ascribe representational content to neural signals too—for example, to say that phasic dopamine firing represents errors of reward prediction. Thus, representational ascriptions enable researchers to connect computational modelling and neural systems with participants' performance in a given cognitive task (cf. Egan 2010, 2014; Coelho Mollo 2020). This connection affords researchers with an "explanatory gloss" (Egan 2014), which allows them to say *what* task participants are trying to solve, and *how*, on the basis of what inferences and reasoning steps, they are trying to solve it.

This way of connecting computational model and neural activity with behaviour in a given task also dissolves the mismatch problem, which, recall, is the problem of accounting for normal psychiatric usage of the term "dysfunction" in the face of possible mismatches between the computational function ascribed to a system and the environment in which the system operates now. If Reinforcement Learning models are environment-neutral (or domain general), and computational psychiatrists using these models ascribe content to their target systems pragmatically, then the mismatch problem does not arise. Experimental tasks just are the environment in which participants operate now. And the specific computational functions ascribed to experimental participants cannot be mismatched, since these ascriptions depend on the degree of empirical adequacy of alternative computational models in capturing participants' data in the task.

Psychiatrists choose experimental tasks that are relevant to evaluate different dimensions of psychiatric illnesses—for example, they use reversal learning tasks to assay belief updating and cognitive control. Based on the task of interest and on the computational functions ascribed to participants, it may turn out that computational psychiatrists ascribe different representations to participants with similar neurophysiological profiles and embedded in similar social environments—Schlagenhauf et al. (2014), for example, ascribed beliefs about the (hidden) state of their reversal learning task only to some of their patients.

These representational ascriptions enable them to clarify in what sense observed performance in a task is impaired, connecting (mis)computation, neural activity and behaviour. Thus, for example, because delusions are species of rigid beliefs, one might expect that schizophrenia patients with delusions would be less likely to flexibly switch their behaviour in a reversal learning task after reversals in reward contingen-

cies in the task. Yet, Schlagenhauf et al.'s (2014) patients exhibited too much switching, and this behavioural profile correlated with reduced ventral striatal activity and higher levels of the severity of their delusions as measured with the PANSS scale. If one appeals to representational ascriptions to make sense of how miscomputations of prediction errors explain these results, then one could hypothesise that delusions, hallucinations and other symptoms of schizophrenia are all "expressions of the same core pathology: namely, an aberrant *encoding* of the precision" of prediction errors. Many symptoms of schizophrenia, that is, would amount to dysfunctions in neural computations involving representations of uncertainty (Adams et al. 2013, 1).

Though perspectival, these representational ascriptions need not be arbitrary or untestable. The content of dopamine activity is generally understood as a reward prediction error (Schultz, Dayan, & Montague 1997). But this ascription is now contested (Colombo 2014b), and will be probably revised, as recent computational and neuroscientific results indicate that dopamine activity encodes dimensions of an error in prediction unrelated to reward (Langdon et al. 2018). While other researchers believe that dopamine activity represents the precision of a prediction error (Adams et al. 2013), different representational ascriptions motivate further testing of alternative computational models of a given task formulated in different modelling frameworks. Results of these tests will help researchers find more adequate explanations of psychiatric phenomena and targets for more effective treatment.

In summary, the mismatch problem does not arise if we understand miscomputation as *p-miscomputation*, and representational ascriptions pragmatically. Representational ascriptions enable computational psychiatrists to link computational and neural results, with the behaviour to be ex-

plained in a given task. While representational ascriptions are pragmatic, they are not arbitrary. They are based on a natural, common, pre-formal understanding of a given task, and of the computational models for that task. While revisable, computational psychiatrists' representational ascriptions remain warranted to the extent they contribute to further explanatory and practical purposes psychiatrists care about.

5. Conclusion

One of the aims of existing accounts of physical computation is to do justice to actual practices in the computational sciences. In this paper, I focused on central modelling practices in computational psychiatry. I considered a perspectival and pragmatist explication, which I called *p-miscomputation*, and a purely mechanistic explication I called *m-miscomputation*. I argued that, compared to *m-miscomputation*, *p-miscomputation* is a better thing to mean by terms like "aberrant prediction error" or "aberrant precision" in the specific research context of Reinforcement Learning and Bayesian modelling for the purpose of explaining clinically relevant phenomena like impaired learning in schizophrenia.

Perhaps, mechanistic accounts as encapsulated in *m-miscomputation* better comport with successful practices in psychiatry grounded in connectionist (e.g., Cohen & Servan-Schreiber 1992), dynamicist (Globus & Arpaia 1994; Durstewitz, Huys, & Koppe, 2020), or network approaches to computational modelling (Wang & Krystal 2014), of which I said nothing here. But, if the point I made in this paper is right, then ideas from some prominent mechanistic accounts of computation, ideas about how to determine computational functions and what role representation should play in computation, are detrimental to achieve the aim of doing justice to actual practices in the computational sciences. An explica-

tion of (*mis*)*computation* grounded in a perspectival pragmatism will be more descriptively adequate and practically fruitful.

Acknowledgements

I am grateful to Andreas Heinz, J. Brendan Ritchie, Corey J. Maley, Dimitri Coelho Mollo, Joe Dewhurst, Nir Fresco, and an anonymous reviewer for their generous comments on previous versions of this paper. This work was supported by the Alexander von Humboldt Foundation through a Humboldt Research Fellowship for Experienced Researchers at the Department of Psychiatry and Psychotherapy, at the Charité University Clinic in Berlin.

References

Adams, R. A., Huys, Q. J., & Roiser, J. P. (2016). Computational psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*, *87*(1), 53-63.

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in psychiatry*, *4*, 47. https://doi.org/10.3389/fpsyt.2013.00047

Ahmed, S. H. , Graupner, M., & Gutkin, B. (2009). Computational approaches to the neurobiology of drug addiction. *Pharmacopsychiatry*, 42(Suppl.1), S144–S152.

Alcaro, A., Huber, R., & Panksepp, J. (2007). Behavioral functions of the mesolimbic dopaminergic system: an affective neuroethological perspective. *Brain research reviews*, *56*(2), 283-321.

Brugger, S. & Broome, M. (2019). Computational psychiatry. In Sprevak, M., & Colombo, M. (eds.). *Routledge Handbook of the Computational Mind*. Routledge, pp. 468-484.

Churchland, P. S. & Sejnowski, T. J. (1992). *The Computational Brain*. Cambridge, MA: MIT Press.

Coehlo Mollo, D. (2020). Content Pragmatism Defended. *Topoi*, *39*, 103–113.

Coelho Mollo, D. (2019). Are There Teleological Functions to Compute?. *Philosophy of Science*, 86, 431-452.

Cohen, J.D. & Servan-Schreiber, D. (1992). Context, cortex and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, 99, 45–77.

Colombo, M. (2019). Learning and reasoning. In M. Sprevak & M. Colombo (Eds.) *The Routledge Handbook of the Computational Mind* (pp. 381-396). New York: Routledge.

Colombo, M. (2014a). For a Few Neurons More: Tractability and Neurally Informed Economic Modelling. *The British Journal for the Philosophy of Science*, *66*(4), 713-736.

Colombo, M. (2014b). Deep and beautiful. The reward prediction error hypothesis of dopamine. *Studies in history and philosophy of science part C: Studies in history and philosophy of biological and biomedical sciences*, *45*, 57-67.

Colombo, M. & Heinz, A. (2019). Explanatory integration, computational phenotypes and dimensional psychiatry. The case of alcohol use disorder. *Theory and Psychology*. https://doi.org/10.1177/0959354319867392

Corlett, P. R., Murray, G. K., Honey, G. D., Aitken, M. R., Shanks, D. R., Robbins, T. W., ... & Fletcher, P. C. (2007). Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. *Brain*, *130*(9), 2387-2400.

Deco, G., & Kringelbach, M. L. (2014). Great expectations: using whole-brain computational connectomics for understanding neuropsychiatric disorders. *Neuron*, *84*(5), 892-905.

Dennett, D. C. (1987) *The Intentional Stance*. Cambridge, MA: MIT Press.

Dewhurst, J. (2018a). Individuation without Representation. *The British Journal for the Philosophy of Science*, 69, 103–116.

Dewhurst, J. (2018b). Computing mechanisms without proper functions. *Minds & Machines*, 28, 569-588.

Dewhurst, J. (2014). Mechanistic miscomputation: a reply to Fresco and Primiero. *Philosophy & Technology*, 27(3), 495-498.

Durstewitz, D., Huys, Q. J., & Koppe, G. (2020). Psychiatric illnesses as disorders of network dynamics. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

Egan, F. (2019). The Nature and Function of Content in Computational Models. In M. Sprevak and M. Colombo (eds.), *The Routledge Handbook of the Computational Mind*. Routledge, pp. 247-258.

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, 170, 115–135.

Egan, F. (2010). Computational models: A modest role for content. *Studies in History and Philosophy of Science Part A*, *41*(3), 253-259.

Feeney, E. J., Groman, S. M., Taylor, J. R., & Corlett, P. R. (2017). Explaining delusions: reducing uncertainty through basic and computational neuroscience. *Schizophrenia bulletin*, *43*(2), 263-272.

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48-58.

Fodor, J. A. (1987). *Psychosemantics*, Cambridge, MA: MIT Press.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Fresco, N. (2014). *Physical Computation and Cognitive Science*. Springer.

Fresco, N., & Primiero, G. (2013). Miscomputation. *Philosophy & Technology*, *26*(3), 253-272.

Friston, K. J. (2010). The free-energy principle: a unified brain theory?. *Nature Reviews Neuroscience*, *11*(2), 127-138.

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148-158.

Garson, J. (2019). *What Biological Functions Are and Why They Matter*. Cambridge University Press.

Globus, G. G., & Arpaia, J. P. (1994). Psychiatry and the new dynamics. *Biological Psychiatry*, *35*(5), 352-364.

Gu, X., Eilam-Stock, T., Zhou, T., Anagnostou, E., Kolevzon, A., Soorya, L., ... Fan, J. (2015). Autonomic and brain responses associated with empathy deficits in autism spectrum disorder. *Human Brain Mapping*, 36, 3323–3338.

Howes, O. D., & Kapur, S. (2009). The dopamine hypothesis of schizophrenia: version III—the final common pathway. *Schizophrenia bulletin*, *35*(3), 549-562.

Huys, Q. J., Maia, T.V., & Frank, M.J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, *19*(3), 404-413.

Huys, Q. J., Daw, N. D., & Dayan, P. (2015). Depression: a decision-theoretic analysis. *Annual review of neuroscience*, *38*, 1-23.

Huys, Q. J., Guitart-Masip, M., Dolan, R.J., & Dayan, P. (2015). Decision-theoretic psychiatry. *Clinical Psychological Science*, *3*(3), 400-421.

Huys, Q. J., Moutoussis, M., & Williams, J. (2011). Are computational models of any use to psychiatry?. *Neural Networks*, *24*(6), 544-551.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia bulletin*, *13*(2), 261-276.

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *science*, *321*(5890), 806-810.

Kurth-Nelson, Z., O'Doherty, J., Barch, D., Deneve, S., Durstewitz, D., Frank, M., & Tost, H. (2016). Computational approaches for studying mechanisms of psychiatric disorders. In A. D. Redish & J. A. Gordon (Eds.), *Computational psychiatry: New perspectives on mental illness* (pp. 77-99). Cambride, MA: MIT Press.

Langdon, A. J., Sharpe, M. J., Schoenbaum, G., & Niv, Y. (2018). Model-based predictions for dopamine. *Current Opinion in Neurobiology*, *49*, 1-7.

Lawson, R. P., Mathys, C., & Rees, G. (2017). Adults with autism overestimate the volatility of the sensory environment. *Nature Neuroscience*, *20*(9), 1293-1299.

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154-162.

Miłkowski, M. (2013). *Explaining the Computational Mind*, Cambridge, MA: MIT Press.

Millikan, R. G. (1984). *Language, thought, and other biological categories*. Cambridge: MIT Press.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, *16*(1), 72-80.

Neander, K. (1991). Functions as selected effects: The conceptual analyst's defense. *Philosophy of science*, *58*(2), 168-184.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139-154.

O'Connell, L. A., & Hofmann, H. A. (2011). The vertebrate mesolimbic reward system and social behavior network: a comparative synthesis. *Journal of Comparative Neurology*, *519*(18), 3599-3639.

Pani, L. (2000). Is there an evolutionary mismatch between the normal physiology of the human dopaminergic system and current environmental conditions in industrialized countries?. *Molecular Psychiatry*, *5*(5), 467-475.

Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford: Oxford University Press.

Rescorla, M. (2014). A theory of computational implementation. *Synthese*, 191, 1277-1307.

Schlagenhauf, F., Huys, Q. J., Deserno, L., Rapp, M. A., Beck, A., Heinze, H. J., & Heinz, A. (2014). Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. *NeuroImage*. 89, 171–180.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

Schweizer, P. (2019). Computation in Physical Systems: A Normative Mapping Account. In *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*. Springer, pp. 27-47.

Shagrir, O. (2018). In defense of the semantic view of computation. *Synthese*. https://doi.org/10.1007/s11229-018-01921-z

Sprevak, M. (2013). Fictionalism about neural representations. *The Monist*, *96*(4), 539-560.

Sprevak, M. (2010). Computation, individuation, and the received view on representation. *Studies in History and Philosophy of Science Part A*, *41*(3), 260-270.

Sutton, R.S. & Barto, A.G. (2018). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tsou, J. Y. (2012). Intervention, causal reasoning, and the neurobiology of mental disorders: Pharmacological drugs as experimental instruments. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(2), 542-551.

Tucker, C. (2018). How to Explain Miscomputation. *Philosophers' Imprint*, 18 (24), 1-17.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 59(236), 433–460.

Turner, R. (2011). Specification. *Minds and Machines*, 21(2), 135–152.

Wang, X. J., & Krystal, J. H. (2014). Computational psychiatry. *Neuron*, *84*(3), 638-654.